Evidence and Design Implications Required to Support Comparability Claims

Richard M. Luecht

The University of North Carolina at Greensboro

Wayne J. Camara

The College Board

The PARCC assessment system is being designed to address a seemingly ambitious set of goals, including reliably and validly assessing the full compendium of the *Common Core State Standards* (CCSS) for Mathematics and English Language Arts (ELA)/Literacy across grades, years, and states for purposes measuring growth and other aspects of accountability, evaluating student readiness for college and careers, and providing limited diagnostics to teachers and students. PARCC has established the following comparability goals: (a) to maintain within-state comparability of scores across schools and districts; (b) to facilitate cross-state comparisons (at least within each of the consortia); (c) to measure growth for accountability purposes across grades; and (d) to evaluate changes or trends from year-to-year[1]. Arguably, the within-state comparison goals are being met to some degree by existing state assessments, including various aspects of accountability and growth across grades and years. However, there has been no success in establishing comparability across states.

The PARCC assessment system is therefore being designed to provide some level of inference about comparability of scores across states by utilizing the CCSS and providing common assessments tasks across states in the consortia. However, the exact way in which common assessments can and will be implemented is yet to be determined. What is clear is that legitimate comparisons across grades, years, and states will require a careful consideration the types of linkages that can be established between examinees and test forms, test design and development practices employed—including test assembly specifications, quality control, modes of administration, standardization of the conditions of measurement, sampling designs and controls, scoring procedures, and the nature and extent of inferences desired with respect to individual student scores and aggregate scores for schools, districts, or states. This paper lays out some of those comparability considerations and several optional test linking designs that would help ensure that

[1] It is our understanding that across-consortia comparisons are not being considered, at present. However, that could change in the future.

required empirical conditions and assumptions necessary for making various types of comparability inferences are actually met.

Score Linking: Statistical Equating and Concordance for Comparability

Test scores can be computed or estimated in many ways, even for the same test form and certainly for test forms assumed to measure the same construct. For example, we can add the points for all items on the test, possibly weighting some test items more than others, or, we can apply an elaborate psychometric scaling model to provide one of more scores from the item scores or patterns of item scores. A set of test scores can also be aggregated (e.g., summed, averaged, or classified into achievement levels and counted) in different ways. We might apply different selection or sampling criteria, use weights, or even sophisticated multilevel statistical models to directly estimate effects for different student groups of interest. Are those individual scores or aggregations comparable? What if <u>different</u> test forms are used for different groups being compared or different states? What about using different item types or administering some tests on a computer and others in paper-and-pencil format? Can any of the scores be made comparable?

At the most basic level, a test score is a number on some scale. From that scale, we typically want to make inferences about a construct—that is an ordered set of expectations or claims about performance. However, in order to compare two or more test scores, we need to ask a basic question. Are the constructs underlying those scores the same, similar, or different? The measurement literature often suggests a basic duality as to what is being measured: the same construct versus different constructs. However, there are degrees of sameness[2]. Two mathematics tests are not necessarily measuring the same construct, especially across different grades.

A score scale is effectively the operational instantiation of a particular construct or mixture of constructs. Score scales may differ in the nature of the phenomena being measured or merely reflect differences in the choice of units. If we measure length in inches or centimeters, the scale is still the same—the units of length merely differ. There are convenient transformations to convert metric lengths to/from U. S. lengths. Test scores are more complex in that we do not have established, convenient scales that measure constructs like mathematics or reading comprehension. Because two tests may use the same name, share a common content blueprint or otherwise claim to measure the same construct by no means implies that the test scores actually function in the same way—statistically speaking—or produce comparable score interpretations. This is one of fundamental dilemmas in assessment—to develop a *scale* that consistently and validly

[2] For example, forms may differ in terms of content coverage, depth of coverage, cognitive demand, response processes required by students, difficulty, sequencing, length, instructions and numerous other features.

measures a particular construct so that relevant comparisons, decisions, and other interpretations can be made.

Test equating is a well-known statistical process for adjusting test scores by placing them on a common scale (Kolen & Brennan, 1995; Holland & Rubin, 1982). Equating adjusts for minor differences in difficulty among two or more test forms built to the same content and other statistical specifications, placing the test scores on a common metric or scale[3]. When test equating works well, any desired comparability among test takers (or aggregated scores for schools, districts or states) is straight-forward in much the same way that a measure of the length of one table in inches and a second table in centimeters would be directly comparable after transforming from U. S. to metric length or vice versa. Consider that, if defensible equating of test scores across grades, years and states were possible, we could legitimately compare James' (who lives in Massachusetts) Grade 8 mathematics score in 2010 to Monica's (who lives in Georgia) Grade 8 mathematics score in 2011. If both students received a score of 90, could we conclude that Monica knew as much math as James?

Ultimately, test equating—that is, putting test scores on a common metric or scale to facilitate comparability—requires that a somewhat rigorous set of measurement conditions and assumptions be met. Equating leads to what is sometimes referred to as score *interchangeability* (Brennan & Kolen, 1995; Holland & Dorans, 2006). After equating, it ought to be a matter of indifference to students, teachers, administrators or policy makers as to which form of the same test or which items each examinee sees. The scores for examinees at the same level of proficiency are interchangeable *because they are on a common scale*.

Unfortunately, the extent to which scores actually are interchangeable is often perceived in fundamentally different ways by measurement experts and other stakeholders in educational assessment (the lay public, students, teachers, educational administrators and policy makers). Measurement experts understand that successful equating requires, at the very least, clear *linkages* or *connections* among the scores being equated: either common (shared) test items or common examinees taking the test form. And, even with common items or examinees, equating only works well when the constructs being measured are verified as being the same[4]. For example, the ACT, the SAT and many state testing programs produce multiple test forms for administration within a year and across years which are equated and produce statistically comparable scores. However, equating is

[3] Item response theory (IRT) uses a combination of item calibration and equating procedures to essentially accomplish the same goal—determining examinees' scores on a common scale.

[4] The term *battery scaling* has been used to describe statistical adjustments that approximately equalize the score distributions for two or more tests in a common reference population, even though the tests may measure different constructs (Angoff, 1971).

almost always restricted to multiple forms of the same test. In contrast, policy makers and member of the lay public may often argue that reading comprehension is reading comprehension regardless of which test is used or how the scores are computed or aggregated. The rather naïve assumption is that the same numerical score (scale score, percentile, etc.) means the same thing for all students, just because the test labels or content are similar. In other words, policy makers and administrators often want to interpret all test scores as being on a common scale, simply by declaration or logic. However, scores from reading comprehension tests in different states, as well as scores on the ACT and the SAT scales, do not meet the requirements of equating. As the simple example above hopefully demonstrates, that interpretation of test scores as being completely comparable—with the same scores being interchangeable from the perspective of decision making and interpretation—would only work if the equating paradigm works, and works very well.

It is important to realize that PARCC and the participating states have options regarding the extent and nature of score comparability interpretations that will be possible across states, grades and years. The goal at this point in time is and should be to devise an assessment system that ideally supports equating. If equating is not possible, due to a myriad of potential confounding factors, some level of comparability may still be possible under the broader framework of *score linking*. Holland and Dorans (2006) defined three types of linking: (1) equating; (2) concordance; and (3) prediction. *Equating* can be viewed as offering highest level of comparability—that is, allowing for exchangeable scores and interpretations. *Scale concordance* is essentially a relaxed form of linking to establish score comparability by building a numerical scale on which scores from different tests that measure ideally similar, but not necessarily equivalent constructs, can be used approximately in the same way and given similar interpretations. Concordant scores still require both tests to measure the same construct, although this may be accomplished through with tests having different content and statistical specifications. In addition, a strong statistical relationship between student scores on both tests are required and the relationship between test scores must be invertible (meaning scores from test A can be converted to the score scale for test B and scores from test B can be converted to the score scale for test A) (Dorans, Lyu, Pommerich and Houston, 1997). Finally, *prediction* uses statistical regression modeling to project expected performance on a criterion measure, based on one or more predictor tests or other indicators such as grades, resources provided, and demographics. Our discussion focuses primarily on equating and concordance.

When states compare scores across school districts, when student performance on state end-of-grade (EOG) or end-of-course (EOC) examinations in various states are compared to National Assessment of Educational Progress (NAEP) scores, or when U.S. test scores are compared to international samples, we are not talking about equated test scores or common scales in a strict measurement sense. In fact, the National Research Council's Committee on Equivalency and Linkage of Educational Tests, Board on Testing and Assessment (1998) concluded that it was not technically feasible to even try to link the many of currently available commercial and state achievement tests to one another.

In contrast to the NRC conclusions, PARCC and the participating states do anticipate establishing solid linkages across test forms by using a common set of test forms across years and grades with required item and task linkages. To the extent that these types of strong linkages are actually implemented and adequately maintained, statistical equating may indeed be possible. However, there are many factors that can undermine the quality of the linkages and limit the types of inferences that might be possible in practice. Therefore, in this paper, we attempt to go beyond the surface-level requirements for equating (e.g., common test forms, common anchor items) and discuss additional threats to score linking and comparability. Ultimately, our recommendations emphasize the need to exercise strict controls over the entire testing enterprise, with solid quality assurance procedures and ongoing invariance studies that evaluate the veracity of those linkages. If the required linkages breakdown or become unstable we may more likely referring to what the *Standards for Educational and Psychological Testing* (AERA, APA, NCME, 1985) call *scaling to achieve comparability*. This concept has been described by other labels such *linking* (Linn, 1993), *alignment* (Holland & Dorans, 2006), or establishing *concordance* (Pommerich & Dorans, 2004).

Item and Test Linking Design Considerations

Scale linking is largely about making statistical adjustments to scores to ideally put them on a common scale. Item and test design can either facilitate or hinder the adjustment mechanisms used. For example, if test forms are designed without providing common item linkages across the forms, it becomes impossible to dissect differences in performance as owing to differential difficulty of the test forms or differential performance of the examinees taking those forms. However, designing tests to explicitly share items across test forms is not, in and of itself, adequate to ensure that equating can be done. The relationships between the common items and the remaining testing items, the size of the link, and interactions with examinee characteristics can all conspire to limit the quality of linkages. This section describes three aspects of item and test design that are relevant to this issue of linkages for the PARCC Assessment System: (a) the nature and quality of common-item linking for equating; (b) linking by content and test specifications; and (c) linking tests that employ multiple item formats and that possibly measure different constructs.

Nature and Quality of Common-Item Linking

It is important to recognize that linkages between test forms can be established in multiple ways. The most obvious approach is to share intact items among two or more test forms. For traditional equating, these links are built by sharing or reusing a relatively small number of items as an "anchor test". Performance can therefore be compared on the anchor test and projected to the total-test scores. Another convenient type of common-item linking is to reuse IRT-calibrated items from an item bank on the test forms. In this case, the linkage is between the item bank and the individual items that comprise a test form, rather than between any two test forms. A third type of linkage is between individual items and an item family or task model, where the task models or templates based on each task model are calibrated to the item bank scale. Here, there is a hierarchical link between items, the templates, the task models and the bank scale (Luecht, 2008; Luecht, Dallas & Steed, 2010; Shu, et al 2010).

Figure 1 demonstrates a rather conventional anchor test design where the common items are shared between two test forms, A and B. This type of linking design uses performance of the common items as the basis for adjusting the Form B results to the (base) Form A scale, even if examinee samples having different proficiency distributions take each form (i.e., for non-equivalent groups equating). Over time, a chain of linking item sets can be generated where all new forms are linked to previously equated forms.

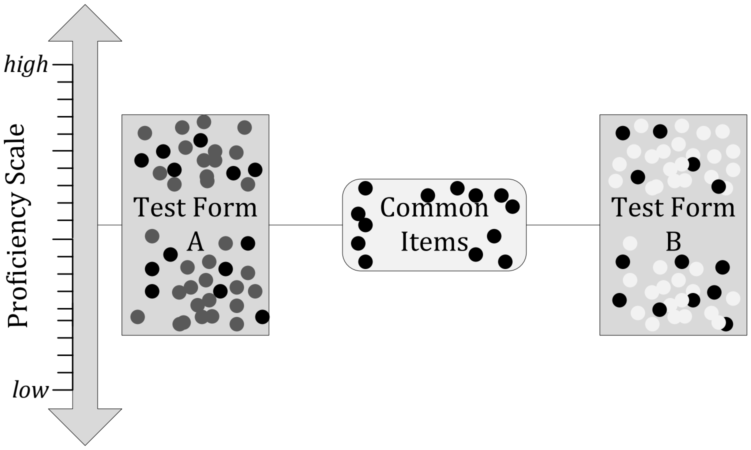

Figure 1. Common Item Linking of Test Forms

Figure 2 depicts a typical IRT linking design. The item bank contains a large collection of *calibrated* items (see, for example, Lord, 1980 or Hambleton & Swaminathan, 1985). Calibration determines the difficulty and other statistical characteristics of the items relative to an established proficiency scale—*the bank scale*. The calibrated items can be conveniently used to assemble numerous test forms by employing automated test assembly (van der Linden, 2005) or other item selection mechanisms, including computerized adaptive testing algorithms (van der Linden & Glas, 2010). Because the calibrated items are already "located" relative to the intended proficiency scale, the linkage of new items to that same scale on any number of test forms can be readily accomplished by anchoring the existing items using their bank statistics. Once calibrated, examinees can receive proficiency scores on the underlying bank scale, regardless of which test form they take.

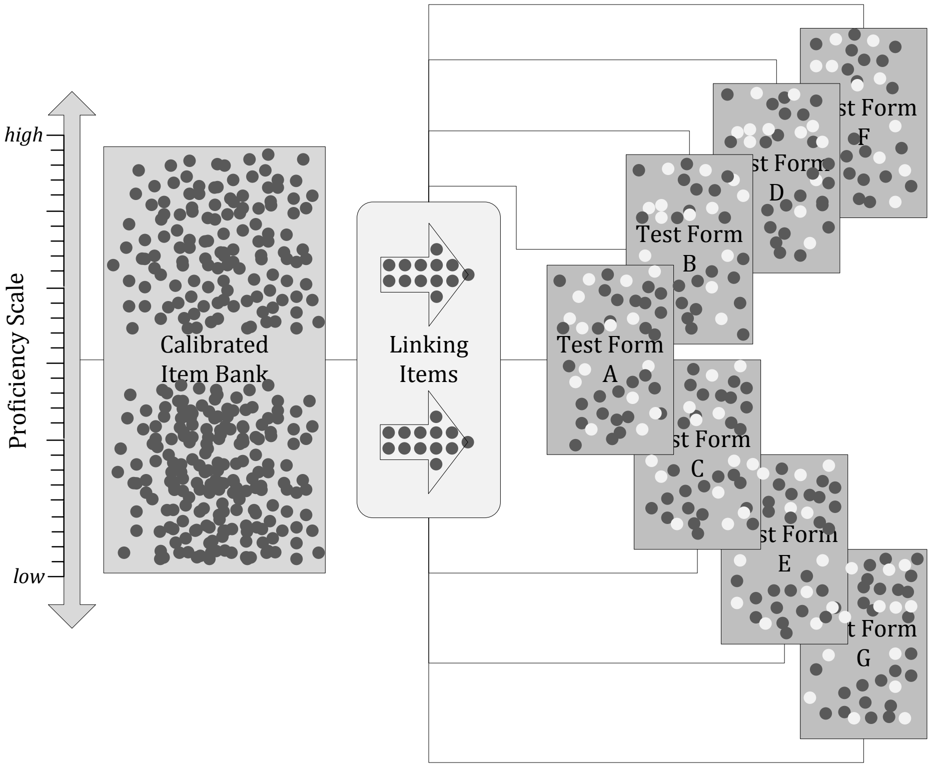

Figure 2. IRT Linking via a Calibrated Item Bank

Figure 3 presents a novel approach to test design and development where templates and items (produced from the templates) are developed to correspond to detailed cognitively based task models. The task models are calibrated—that is, "located"—on the proficiency scale by design and hierarchical IRT calibrations (e.g., Glas & van der Linden, 2007) are used to calibrate the items, templates or, ideally, the task models, themselves. This result is an item production system, where the hierarchical relationships between items, templates and task models provide built-in quality control mechanisms for monitoring the quality of every link in the system[5]. This *assessment engineering* (AE) approach to test design has not been formally implemented for any operational testing programs; however, some promising proof-of-concept research has been successfully

[5] Large variation in the calibrated item statistics usually signals a need to tighten the specifications for a particular template associated with a given task model (see Shu, Burke & Luecht, 2010).

carried out[6] (e.g., Bejar et al, 2003, Lai, Gierl & Alves, 2010; Luecht, Burke & Devore, 2009; Masters & Luecht, 2010).

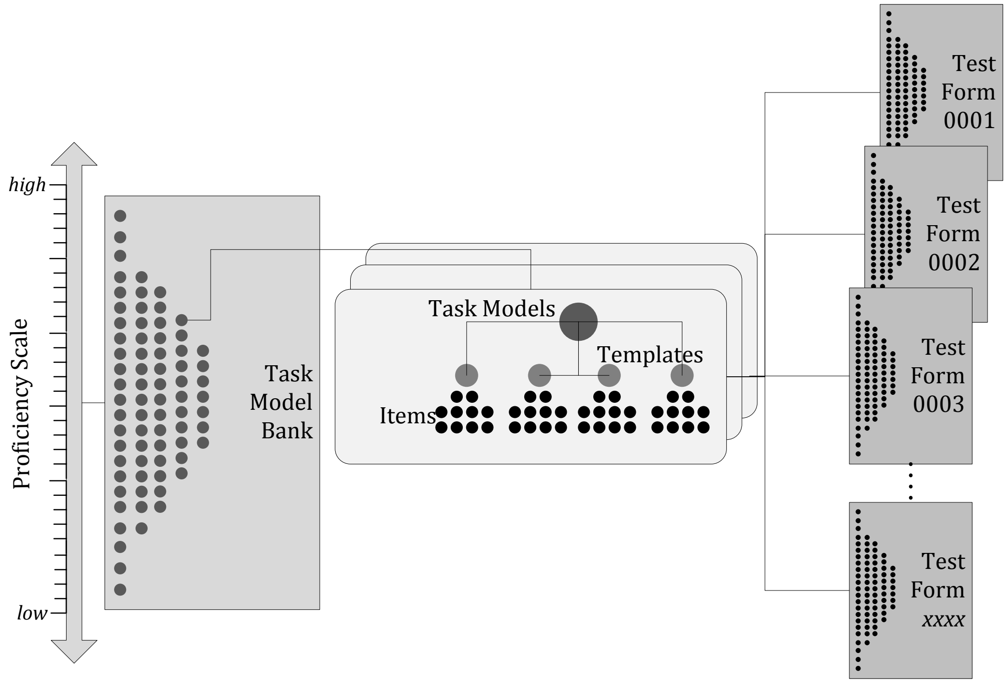

Figure 3. IRT Linking via Calibrated Task Models and Item Templates

Although it is unlikely that a fully functional AE-based test production system could be put in place for the PARCC consortium by the intended implementation date, this approach provides an interesting contrast to traditional direct linking designs based on shared anchor items or reused calibrated items from a bank. At the very least, merely mentioning AE helps to emphasize the strong need for having a principled approach to test design, item writing and test-form assembly—regardless of the linking design ultimately chosen. That is, whether one adopts a more traditional common-item equating design, an IRT calibrated bank framework, or a hierarchical AE linking design, the quality of the item links and linking process, itself, require constant monitoring. Item parameter drift, differential item functioning, and other methods of monitoring the statistical robustness of the linkages need to be implemented as an integral part of the PARCC testing enterprise.

Linking by Content and Test Specifications

Test specifications typically include content requirements and statistical targets for each test form.   Content requirements are often indicated using item counts for various content and cognitive codes.   The test specifications can also include item types, reading

[6] An area related to AE is called *automated item generation*, where item shells or models (i.e., templates) are used to manufacture many variants of a particular item (Haladyna, 2004).

levels, general topics, or almost any type of categorical coding scheme desired. In the context of the *Common Core State Standards* (CCSS), the test specifications would reflect the high priority mathematics and English language arts (ELA)/literacy elements selected by subject-matter experts for inclusion on the assessments—that is, mastery claims and supporting reporting categories that reflect a relative balance of content emphasis for portions of the content standards, degree of scaffolding expected between grades, complexity of texts, etc.. If states were to develop their own tests using the CCSS, would it still be possible to link the scores? The simple answer is "no"—at least not in an equating sense of comparability.

The simple fact is that all content specifications are fallible insofar as: (a) being judgmentally determined by subject-matter experts; (b) somewhat arbitrarily specified at the test level; and (c) usually not demonstrating much relationship to the statistical characteristics of the test, even if the content specifications are exactly met. The test specifications typically also include statistical targets such as average scores, minimum acceptable levels of reliability, or IRT-based test information targets. However, even using similar statistical targets and employing similar test assembly procedures (e.g., automated test assembly algorithms) is insufficient to establish a necessary level of linkage for equating to hold.

Without tangible links—either direct anchor item links between test forms or hierarchical links through item families—testing equating is simply not feasible. This point emphasizes an important caveat for the PARCC assessment system. Implementing the CCSS across the consortium, by itself, is not sufficient to establish the types of linkages needed for equating. Even if states adopt common test assembly practices (content specifications and statistical targets, similar test assembly practices, etc.), formal linkages will still need to be designed into every PARCC test form. If the CCCS are the primary basis for the linkages concordance or prediction might be possible based sole on similarities in test content and assembly practices, but most direct comparability inferences would be severely curtailed and there will be obvious limitations for what policy makers can say about the scores that result from a scale concordance view or a predictive of the PARCC assessment system. At best, concordant scales will allow for *approximate comparisons* among grades, from year-to-year or across states in the consortium. Strong inferences about differences or effects, or absolute comparability claims, need be tempered by the limits of approximation. This is not merely a philosophical argument that can be conveniently ignored by policy makers. We cannot say that concordant end-of-year (EOY) scores are measuring the same thing, even though the same CCSS blueprints are used in test development, the same number units are reported, common standards are used, the interpretive guides more or less say the same thing, and score distributions of percentages of students in various achievement categories that look the similar. Approximations are still approximations.

Fortunately, PARCC is planning to use intact test forms across states, as well as across grades and over time—at least for the through-course assessment (TC_3) and the end-of-year assessments (EOY). It is therefore likely that either IRT-based item banking or common-item, form-to-form equating will be possible at least across states and over time. If solid, common-item links can be designed into the PARCC test forms—ideally reusing intact test forms across all states—it should be feasible to directly compare scores at least within grades[7].

Linking Assessments Employing Different Item Formats and Complex Content

The variation between test scores can be categorized according to three broad sources: (1) variation due to actual differences between students on the proficiency scale of interest—i.e., the intended construct; (2) variation due to random measurement errors; and (3) systematic variation due to assessment methods, auxiliary traits, or other so-called *nuisance factors* that impact students' scores. We expect students to naturally differ in their proficiency scores, so the first type of variation is entirely acceptable from a measurement perspective. By building tests that target measurement precision where it is most needed along a scale, we can also contend with the second type of score variance—ideally minimizing measurement errors through sound test assembly practices. The third type of variation could be more problematic insofar as indirectly impacting the comparability of scores—especially affecting any composite scores (i.e., overall scale scores or performance levels based on composite scores) that in some way depend on constructed response items, performance assessment items, or technology-enhanced items that many involve somewhat memorable content or problem-based scenarios.

There are many legitimate reasons for the popularity and proliferation of multiple-choice (MC) test items. They are relatively inexpensive to design, produce and score, can be generated in sufficient numbers and mixed-and-matched across test forms to reduce exposure risks, and are generally amenable to established item-analysis procedures and IRT-calibration methods (Haladyna, 2004). Critics of MC items often cite their limited potential to measure high-order cognitive skills, instead recommending the use of skill-based performance assessments and constructed-response items ranging from short-answer responses to essays and other integrated tasks with complex response types that require either scoring by require human raters or *intelligent*, automated scoring engines. Beyond the added costs of scoring, one of the recognized problems with using complex

[7] The discussion of multidimensionality in this section is germane to the issue of equating scores across grades, especially when the nature of the construct and scales undergo incremental changes in the nature of what is being measured.

item types is that they tend to memorable and may pose significant *exposure risks* where some students may have actually seen the items before and therefore have a somewhat unfair advantage. For example, writing prompts are usually highly memorable, making it difficult to reuse them over time. Other types of performance assessment items that use recognizable stimuli or exhibits may have similar exposure risks when reused for equating purpose. This presents a serious dilemma for equating. How can equating links be established if these complex item types are not reused?

If these more complex item types were merely measuring the same trait as multiple-choice items, the challenge of equating composite traits might not be too serious. The simple fact is, however, that most constructed-response and complex item types are expressly introduced to measure "something else". To the extent that these types of items do measure somewhat different dimensions than the MC items, or otherwise interact with differential language traits, learning opportunities, or auxiliary traits specific to certain test-taker subgroups (suggesting differential item or test functioning), it may be difficult or even impossible to establish the necessary linkages across the composite scale for equating. This may be true for the PARCC through-course (TC_3) assessments and for the end-of-year (EOY) assessments, considered independently—especially if free-response or technology-enhanced items cannot be conveniently reused as linking items over time. It will almost certainly be a consideration if states want to combine the TC_3 and EOY assessments to form a composite scale score that is assumed to be equated over time.

The fundamental problem can be demonstrated by a simple tri-variate correlation (or covariance) matrix:

$$R = \begin{pmatrix} 1.0 & & \\ r_{MC,CR_1} & 1.0 & \\ r_{MC,CR_2} & r_{CR_1,CR_2} & 1.0 \end{pmatrix}$$

where the context is a test comprised of both MC and constructed-response (CR) or performance-based item types. There are three potential proficiency traits: (a) the primary knowledge trait measured by the MC items; (b) a particular collection of constructed-response tasks measuring a performance-based trait in year #1 (denoted CR_1); and (c) a new collection of constructed-response tasks measuring a related performance trait in year #2. (denoted CR_2). The MC dimension is assumed to remain relative constant over years, with equating carried out by common-item links between forms or IRT calibrations to a bank scale.

Even if the two CR traits were demonstrated to be invariant across years[8], the relationship between each CR trait and the MC trait would need to be virtually identical (i.e., $r_{MC,CR_1} = r_{MC,CR_2}$) for a composite of the MC and CR traits to have any chance of holding up under an equating paradigm based only on the MC items. If r_{CR_1,CR_2} is less than one, the implication is that CR_1 and CR_2 are different across years, perhaps due to differences in topics, stimuli, scoring rubrics or other design-related characteristics of the CR items. Since we cannot estimate r_{CR_1,CR_2} (see footnote), a composite scale based on MC+CR_1 could be substantial different than a composite scale based on MC+CR_2, even after equating the MC scales across years.

The point is that the only legitimate way to ensure that equating will work for composite scales (e.g., TC and EOY) will be to establish strong equating links across all item types. The method of equating/calibration is not the issue. Rather, it is about the confounding of the dimensionality of the scales over time and the potential inability of the item or task linkages to maintain a consistent composite scale. It should be noted that a similar dimensionality problem arises for vertical scales where the nature of the composite scale changes across grades.

Score Concordance as an Option

Although the PARCC Assessment System is being planned to support equating by using intact TC_3 and EOY forms across states and anchor items across forms (or using calibrated IRT item banks), there remains the possibility that there are too many moving parts and too many potential sources of variation to across states, time and grades to assume that the constructs and scales are anything more than approximations supported by score concordance rather than equated scores. Fortunately, even approximations are useful. As the distinguished mathematical statistician, John Tukey, wrote:

> "*Far better an approximate answer to the right question, which is often vague, than an exact answer to the wrong question, which can always be made precise*." (p. 13)

The reality is that the Common Core State Standards do not specify in any concrete way the ordered skills, knowledge, and conditions of measurement needed to define the requisite mathematics and ELA/literacy constructs—possibly showing progressions across grades. To permit equating, subtle and not so subtle variations in the actual design and development of assessments deployed in each of the states would need to be significantly reduced or even eliminated— especially with respect to different "innovative" item types used for the through-course (TC) assessments and manner in which those TC assessments

[8] This would be impossible to verify under this scenario because the CR items are not reused over years and no common-person linkages exist for estimating r_{CR_1,CR_2} .

are combined with the end-of-year (EOY) assessments. Even in isolation, any potential variations across states in test design, delivery, specifications, or administration and scoring, would probably be sufficient to raise serious questions about the comparability of the constructs (score scales) across grades, years and states. If multiple variations were considered jointly, the potential for violations of the construct equivalence argument increases by an order of magnitude[9].

Our point is not to recommend against even attempting equating. Rather, there needs to be an alternative plan for comparing scores if equating proves to be untenable. PARCC may need to relax its expectations for the scales in line with what concordance allows, we gain some of the realistic flexibility needed to implement the assessment system. Recognizing the approximate nature of concordant scores, we would further need to constantly recommend against strong inferences that over-generalize or misinterpret the scores and comparisons based on those scores.

However, caution is warranted even with respect to concordance. A potential fallacy is that concordance can be used whenever formal test equating is not viable. This may suggest that since the conditions required for equating cannot be met because of the desire to permit variations across states in the types of items and other major assessment characteristics, , we can scale back the type, quality and number of empirical and conceptual linkages between different TC and EOY examinations created and deployed by each of the states. Score concordance—at least as that phrase is used in this paper—will require serious attempts to ensure comparable sampling designs, data collection methods, choices of models and calibration/statistical equating methods used for concordance, decisions on criterion, management of method variance, choices of item types and associated measurement information used in scoring, test design and development practices—including statistical targets, task models, and content constraints, development of solid concordance links among test forms, standardization, management of security risks/violations, instructions, timing and sequencing of the assessments within an academic calendar, and stable opportunities to learn insofar as training/curriculum design related to the CCSS. Specifically, concordances are population dependent, which means that they are normally computed across common students and if the sample of students taking both tests is not representative of the population, the results will have limited generalizability.

There are also at least three limitations of concordance especially with respect to extreme scores (e.g., near the lower percentiles). First, measurement precision of most test score scales is typically worst near the tails of a score distribution as conditional information drops off. Second, the number of examinees near the either tail of a score distribution is usually small, even with large overall samples, and increases the conditional

[9] One could argue that the forms for all PARCC states will be exactly the same and that they will be administered under exactly the same standardization conditions. In that case, equating seems plausible. Here, we are offering a qualification to allow for conditions of measurement or other factors that might cause the tests to not be exactly the same or to not be administered under exactly the same conditions.

sampling error. This would seem to introduce a potentially serious problem for frequency-based or equipercentile concordance/equating methods. Finally, if pre- or post-smoothing are applied, most localized statistical smoothers—while working well to reduce equating errors in the mid-range of a score scale (Kolen, 1991; Hanson et al, 1994)—may become less stable near the tails. This smoothing stability phenomena near the tails has not been well researched in the equating literature (Kolen & Brennan, 2004) but is reasonably well documented in the graphical smoothing literature for localized smoothers such as cubic splines, LOESS, and distance-weighted least squares (see, for example, Wilkerson, 1999). The instability is further not isolated to the choice of smoother, but may also be affected by choices of bandwidth and kernel functions.   Considered together, measurement errors, sampling errors and numerical smoothing errors could actually be compound near the tails and seriously impact the estimated concordances in those regions of the score scale.

Whereas equating strives to achieve fully exchangeable scores and scaling matches the distribution of scores, concordances provide exchangeable scores within the limits of the precision of each test and the relationship (typically measured through a correlation) between tests. The concordance is also sample-dependent as noted above. Other scaling or linking methodologies exist which have less rigorous requirements and assumptions placed on the comparability of test forms and items, but these methods also offer less rigor in comparing scores. For example, a third type of correspondence is a predictive model which attempts to predict scores on one test given prior performance on a second test. Unlike equating and scaling methods, prediction relationships are not symmetric and relationships exist in one direction (Dorans et al., 1997). Predictive relationships typically produce expectancy tables which illustrate the predicted score on test B given a specific score on test A. A second table is required to predict score on test A given scores on test B and the score ranges and precision of predicted estimates will likely vary across the scale and tables. This method is appropriate when there is only a moderate correlation between tests and there are more significant differences between samples taking each test and the test specifications.

Some Recommendations

This paper takes a pragmatic view on comparability and presents equating as the *gold standard* for score comparability across states, time and grades. Several equating designs were described with common-item anchoring or IRT item bank calibrations probably being the more practical methods of linking the PARCC mathematics and ELA/literacy scales. Some of the challenges and threats to linking are presented with the caution that, ultimately, equating many be difficult to justify for certain types of comparisons or under certain types of scenarios (e.g., not reusing constructed-response or performance assessment items over time).   If the linkages break down, it is probably unrealistic to expect that a formal equating paradigm will work for the PARCC assessments. However, we do not recommend scaling back any of the design plans or resources merely because interchangeable scores across grades, years and states are not viable for this assessment system. We do recommend that the following conditions be given strong consideration.

First, well-articulated, cognitively-based constructs, based on the CCSS, should be developed that lay out the ordered claims and evidence requirements for mathematics and ELA/literacy performance within each grade. This approach to construct specification is consistent with evidence-centered design (Mislevy, Steinburg & Almond, 1999; Mislevy, Almond & Lukas, 2003; Mislevy, Wilson, Ercikan & Chudowsky, 2003) and other approaches to principled assessment design (e.g., Luecht, 2008, 2011; Luecht, Dallas & Steed, 2010). Any “highlighted domains” should be reflected in these construct specifications. Evidence centered design (ECD) approaches may offer some advantages over conventional item design and test specifications because such new design approaches prioritize more explicit connections between item from task models which are directly derived from evidence. Such task models and approaches can better control for content, cognitive demand and statistical properties and result in forms which are more parallel and statistical comparable (Hendrickson, Huff, Luecht, 2010). This approach, in turn may reduce the burden placed on scaling to produce comparable scores and also permit more efficiency in form development over time. Of course, ECD will not compensate for variations that would be introduced if states within the consortium choose to select different items or item types for their assessments.

Second, the CCSS need to be framed as cognitively based task models. This approach effectively replaces content outlines, skill categories, and item type specifications with an integrated blueprint for item development that incorporates the cognitive skill, knowledge and conditions of measurement requirements for an entire family of items that would provide targeted measurement evidence along the construct.

Third, a common item banking system and test assembly system would be extremely useful. To the extent that common items (or at least templates for items) could be developed, piloted, calibrated, and shared across the consortium, the quality of the concordance linkages would be vastly improved. In addition, a common item banking system, with a greater degree of shared items and less variation across states, would provide improved quality control across the consortium and possible ways to evaluate construct shifts in a research mode. A common test assembly system would likely encourage greater standardization of the test specification and assembly practices across states. Costs of implementing a large-scale automated test assembly (ATA) system would greatly improve the capability for states to develop on-demand assessments on a common, concordant scale, for dealing with different assessment demands and logistical complexities within each state. To the extent that every PARCC state uses exactly the same test forms constructed by a single entity, this recommendation may be unnecessary.

However, if states want to incorporate additional assessment components (e.g., their own versions of the first or second through-course assessments), having a common system would be strongly advised to encourage efficiency, portability and sharing across states.

Fourth, a comprehensive plan for experimental item tryout and field testing should be developed. This plan should recognize that experimental studies involving prototype items may not need to occur as part of operational testing. Creative solutions for low-cost prototype tryouts should be sought. Pretesting using motivated student samples should be performed with any items being used operationally on any PARCC assessments. As discussed above, test assembly from a common system, especially if items were calibrated in a joint item banks, would improve the concordance relationships among scores and also improve the consistency of TC and EOY test forms constructed within each of the states. If within-state, calibrated item banks are the only option across-state linkages will be more *ad hoc* in nature and reduce somewhat the types of comparability inferences that can be drawn.

Fifth, the consortium needs to generate a clear set of standardization protocols, including accommodations for special populations, handling of non-standard administrations, clearly specified and enforced security protocols, and dual-mode[10] (paper-and-pencil versus computer-based testing) administration protocols.

Finally, a comprehensive equating/calibration framework need to be developed that optimizes the linkages of common items (and possibly common examinees) and minimizes potential contaminating factors. The choice of models should derive from the interpretative argument that will also constitute the validity framework for the PARCC assessment (Kane, 2006). As noted earlier, the types of inferences about comparability across grades, years and states are an important determinant of the chosen scaling methodology. Choices of models (e.g., classical test, IRT model choices), sampling design requirements, equating/calibration methods, and criterion for deciding on the final results should be included in this specification.

Acknowledgments

The authors wish to thank the PARCC Technical Advisory Committee, and in particular, Ron Hambleton, Derek Briggs, and Bob Brennan for their many insightful and highly useful comments on an earlier draft of this paper.

[10] Dual-mode administrations—especially for computer-based tests using technology enhanced items that cannot be deployed on paper-and-pencil tests—could potentially alter the constructs being measured under each model. While comparability for dual-mode tests is not within the scope of this paper, it does warrant some mention.

References

Angoff, W. H. (1971). Scales, norms and equivalent scores. In R. L. Thorndike (Ed.). *Educational measurement*, (2nd edition, pp. 508-600). Washington, DC: American Council on Education.

Bejar, I. I., Lawless, R. R., Morley, M. E., Wagner, M. E., Bennett, R. E. & Revuelta, J. (2003). A feasibility study of on-the-fly item generation in adaptive testing. *The Journal of Technology, Learning, and Assessment, 2(3)*. (URL online: http://ejournals.bc.edu/ojs/index.php/jtla/article/download/1663/1505)

Committee on Equivalency and Linkage of Educational Tests, Board on Testing and Assessment, National Research Council. (1998). *Equivalency and Linkage of Educational Tests (Interim Report)*. The National Academies Press (URL online: http://www.nap.edu/catalog.php?record_id=9525)

Dorans, N. J., Lyu, C.F., Pommerich, M., & Houston, W.M. (1997). Concordance between ACT assessment and recentered SAT I Sum Scores. College and University, 73(2), 24-34.

Glas, C. A. W. & van der Linden, W. J. (2003). Computerized adaptive testing with item cloning. *Applied Psychological Measurement, 27(4)*, 247–261.

Haladyna, T. M. (2004). *Developing and validating multiple-choice test items (3rd edition)*. Mahwah, NJ: Lawrence Erlbaum Associates.

Hambleton, R. K. & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston: Kluwer.

Hanson, B. A., Zeng, L., & Colton, D. (1994). *A comparison of presmoothing and postsmoothing methods in equipercentile equating*. (ACT Research Report 94-4). Iowa City, IA: American College Testing.

Hendrickson, A., Huff, K. & Luecht, R (2010). Claims, evidence, and achievement level descriptors as a foundation for item design and test specifications. *Applied Measurement in Education, 23*, 358-377.

Holland, P. W. & Rubin, D. B. (1982). *Test equating*. New York: Academic Press.

Holland, P. W. & Dorans, N. J. (2006). Linking and scaling. In R. L. Brennan (Ed.), *Educational measurement* (4th edition, pp. 187-220). Washington: American Council on Education and Praeger.

Jaeger, Richard M. (1996) *Content congruence as a factor in the linking of state assessments to NAEP*. Paper presented at the Council of Chief State School Officers Large-Scale Assessment Conference June 23-26, Phoenix, AZ. University of North Carolina, Greensboro.

Kane, M. (2006). Validation. In R.L. Brennan (Ed.) *Educational measurement* (4th edition, pp. 17-64). Washington: American Council on Education and Praeger.

Kolen, M. J. (1991). Smoothing methods for estimating test score distributions. *Journal of Educational Measurement, 28*, 257-282.

Kolen, M. J. & Brennan, R. L. (1995). *Testing equating: methods and practices*. New York: Springer.

Lai, H., Gierl, M. & Alves, C. (2010, April). *Generating items under the assessment engineering framework*. Invited symposium paper presented at the Annual Meeting of the National Council on Measurement in Education, Denver, CO.

Linn, Robert L (1993). Linking results of district assessments. *Applied Measurement in Education*, 6, 83-102.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Mahwah, NJ: Lawrence Erlbaum Associates.

Luecht, R. M. (2008, February). *Assessment engineering*. Session paper at the Assessment Engineering: Moving from Theory to Practice, Coordinated panel presentation at the Annual Meeting of the Association of Test Publishers, Dallas, TX.

Luecht, R. M. (2011, March). *Assessment design and development, version 2.0: From art to engineering*. Invited, closing keynote address at the Annual Meeting of the Association of Test Publishers, Phoenix, AZ.

Luecht, R. M., Burke, M., & Devore, R. (2009, April). *Task modeling of complex computer-based performance exercises*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Diego, CA.

Luecht, R. M., Dallas, A., & Steed, T. (2010, April). *Developing assessment engineering task models: A new way to develop testsSpecifications*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Denver, CO.

Mislevy, R. J. (1992) *Linking educational assessments: Concepts, issues, methods, and prospects*. Princeton, NJ: Educational Testing Service Policy Information Center.

Mislevy, R. J., Almond, R. G. & Lukas, J. F. (2003). *A brief introduction to evidence-centered design*, RR-03-16, July 2003 Princeton, NJ: ETS

Mislevy, R. J., Wilson, M. R., Ercikan, K. & Chudowsky, N. (2003). Psychometric principles in student assessments. In T. Kellaghan & D. L. Stufflebeam (Eds.), *International handbook of educational evaluation*, pp. 489-532. Boston: Kluwer.

NCME Committee on Assessment Policy and Practice (2011, June). *Can the assessment consortia meet the intended comparability goals? Or, what types of comparability goals can be met?* Presentation at the CCSSO NCSA Conference.

Pommerich, M. & Dorans, N. J. (Eds). (2004). Concordance [Special Issue]. *Applied Psychological Measurement, 28(4)*.

Shu, Z., Burke, M. & Luecht, R. (2010, April). *Some quality control results of using a hierarchical Bayesian calibration system for assessment engineering task models, templates, and items*. Invited symposium paper presented at the Annual Meeting of the National Council on Measurement in Education, Denver, CO.

Tukey, J. (1962). The future of data analysis. *Annals of Mathematical Statistics 33 (1)*.

van der Linden, W. J. (2005). *Linear models for optimal test design*. New York: Springer.

van der Linden, W. J. & Glas, C. A. W. (2010). *Elements of adaptive testing*. New York: Springer.

Wilkerson, L. (1999). *The Grammar of Graphics*. New York: Springer.